# CSP554—Big Data Technologies

## Assignment #9

## Name : Oduri Sai ram

## ID : A20522183

**Exercise 1) 5 points**

Read the article “Real-time stream processing for Big Data” available on the blackboard in the ‘Articles’ section and then answer the following questions:

1. (1.25 points) What is the Kappa architecture and how does it differ from the lambda architecture?

**Ans.** The batch layer does not do data precomputation on a regular basis. The stream processing system handles all computing. Recomputation is only conducted when the business logic changes by replaying previous data. For this, the Kappa Architecture employs a strong stream processor capable of dealing with data at a quicker rate than the data itself.

Lambda architecture, on the other hand, is a system made up of three layers: batch, speed, and serving. It addresses both the Volume and Velocity challenges of big data. It has a batch-oriented system as well as a real-time system.

1. (1.25 points) What are the advantages and drawbacks of pure streaming versus micro-batch real-time processing systems?

**Ans.** Storm and Samza are pure stream-oriented systems with very low latency and moderately high per-item costs; batch-oriented systems, on the other hand, provide unrivaled resource efficiency at the expense of excessively high latency for real-time applications. In microbatch real-time processing systems, data is buffered and processed in batches. It increases efficiency while decreasing the amount of time an individual item spends in the data flow. Storm Trident and Spark Streaming are two such instances.

1. (1.25 points) In few sentences describe the data processing pipeline in Storm.

**Ans.** In Storm, a topology is a data pipeline or application. Spouts are nodes that accept data and thereby initiate the data flow in the topology. Spouts send tuples to bolts, which do processing, write data to external storage, and may send tuples downstream. Storm groups govern data flow between nodes. Storm defaults to round-robin distribution of spouts and bolts, but the scheduler may be adjusted to suit scenarios when a certain processing step must be done on a single node. Storm provides at least one processing option via an acknowledgment feature that keeps track of the processing state of every single tuple as it goes through the topology.

1. (1.25 points) How does Spark streaming shift the Spark batch processing approach to work on real-time data streams?

**Ans.** Spark streaming pushes the batch-processing approach towards real-time needs by chunking the stream of incoming data items into tiny batches, converting them to DDs, and processing them as normal. It also handles data flow and dissemination automatically.

**Exercise 2**

Extract Kafka Package

A computer screen with white text

Description automatically generated

Install Kafka

A computer screen with white text

Description automatically generated

Command -- bin/zookeeper-server-start.sh config/zookeeper.properties

A screen shot of a computer screen

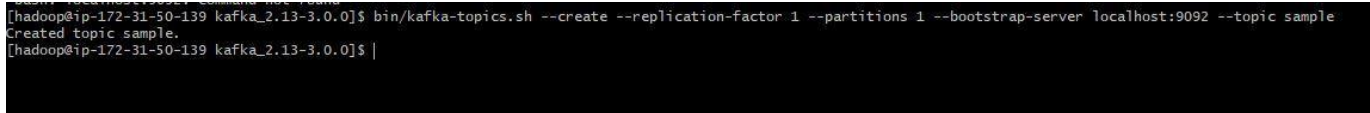
Description automatically generated

A screen shot of a computer screen

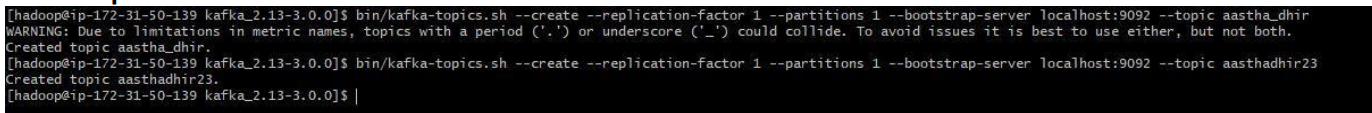
Description automatically generated

At Producer terminal

Sample topic creation



More topics created



Vim put.py

A screenshot of a computer

Description automatically generated

A black background with white text

Description automatically generated

At Consumer terminal

A computer screen with text and images

Description automatically generated

A black background with white text

Description automatically generated